GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment

Year: 2023 | Citations: 441 | Authors: Dhruba Ghosh, H. Hajishirzi, Ludwig Schmidt

Abstract

Recent breakthroughs in diffusion models, multimodal pretraining, and efficient finetuning have led to an explosion of text-to-image generative models. Given human evaluation is expensive and difficult to scale, automated methods are critical for evaluating the increasingly large number of new models. However, most current automated evaluation metrics like FID or CLIPScore only offer a holistic measure of image quality or image-text alignment, and are unsuited for fine-grained or instance-level analysis. In this paper, we introduce GenEval, an object-focused framework to evaluate compositional image properties such as object co-occurrence, position, count, and color. We show that current object detection models can be leveraged to evaluate text-to-image models on a variety of generation tasks with strong human agreement, and that other discriminative vision models can be linked to this pipeline to further verify properties like object color. We then evaluate several open-source text-to-image models and analyze their relative generative capabilities on our benchmark. We find that recent models demonstrate significant improvement on these tasks, though they are still lacking in complex capabilities such as spatial relations and attribute binding. Finally, we demonstrate how GenEval might be used to help discover existing failure modes, in order to inform development of the next generation of text-to-image models. Our code to run the GenEval framework is publicly available at https://github.com/djghosh13/geneval.